The contents of this masked dataset is based on preset decision making by the coaching staff and quarterback of the team. The documentation of the dataset and the overall challenge of the task gave insight that the amount of yards an offense gains, can be influenced by features within the dataset. This made choosing the Random Forest Regression model a comfortable choice for this predictive learning project.

The dataset was loaded into a pandas DataFrame in Jupyter's Python environment then taken through some exploratory analysis of the content types of each column along with the number of rows and columns. The data was then divided into two subsets: one with the yards gained missing and one where yards gained are known. For training and testing, the gain column is converted to a float content type for the yard's known subset. Feature sets were play type, down, and distance for each subset with play type being one-hot coded to allow for binary features for each play type category. The yard's known subset is then split into a training and testing set to evaluate the performance of the Random Forest model.

After training the Random Forest Regressor on the training set, the model is then evaluated for predictive accuracy with Mean Squared Error (MSE). The model met an ideal predictive accuracy result and then predictions on the missing gain subset commenced. There was worry that the scikit-learn package would double the prediction for the csv file so the final two steps, before creating the csv deliverable, was to check the array lengths of the missing play ids and the missing predictions then create a trimmed version on the prediction dataset to match the play ids if necessary.